Problem Chosen	2026 MCM/ICM Summary Sheet	Team Control Number
C		2631870

Can Skill Survive Popularity? Inverse Inference of Latent Votes and the ACE Scoring System

Summary

The scoring mechanism in *Dancing with the Stars* creates a "Superstar Effect," where popularity often overshadows technical merit. To resolve this, we propose a framework integrating Inverse Inference, Causal Inference, and Control Theory.

For Task 1, we employed the **Maximum Entropy Inverse Inference (MEII)** model to reconstruct latent fan votes via **SLSQP optimization**. Results identify "Fan Loyalty" as the primary survival driver (85.3% consistency), validated by robust 95% Prediction Intervals.

For Task 2, we quantified unfairness via an **Institutional Counterfactual Simulation (ICS)** framework. Using **mechanism design theory**, we simulated outcomes under different rules, proving the "Percent Rule" amplifies vote variance compared to the "Rank Rule," creating systemic anomalies **exemplified by the "Bobby Bones Effect."** Our simulation further reveals that the **"Judge Save" mechanism fails to fully correct** this bias. This flaw allows low-skill celebrities to survive **2–3 weeks** longer than warranted, quantified by a **Bias Index rise (0.34 vs. 0.21)**.

For Task 3, we proposed the **Multi-stage Differentiated Voting (MDV) Model** based on **Standardized Differential Regression** to isolate the "Preference Gap" between judges and fans. Analysis quantifies that **Professional Partner capability** drives judges' scores ($\beta = 0.18$) significantly more than fan votes ($\beta = 0.09$), while "Reality TV" backgrounds create a reverse **"Populist Shield"**, necessitating influence rebalancing.

For Task 4, to restore competitive equity, we proposed the **Adaptive Consensus & Evolution System (ACE)** utilizing PID Control Theory.Transforming the static scoring problem into a dynamic control challenge, ACE treats "fairness divergence" as an error signal to regulate judge weights in real-time. Simulations confirm this feedback loop reduces unjustified eliminations by **approx. 85%**, balancing technical merit with audience engagement.

In addition, a **sensitivity analysis** was conducted to systematically validate our framework using **Sobol Indices**. Results confirm that our ACE system maintains stability against extreme voting behaviors and parameter perturbations, demonstrating the model's reliability across diverse contest scenarios.

In conclusion, our Pareto-superior solution transforms an entertainment problem into a control engineering challenge. Crucially, ACE restores fairness by incentivizing "Technical Growth" (the "Cinderella Story") over static popularity, ensuring commercial viability. Ultimately, our framework offers a robust paradigm for **quantifying fairness in hybrid social choice systems**.

Keywords: **Maximum Entropy Inverse Inference**; **Counterfactual Simulation**; **ACE Scoring System**; **PID Control Theory**; **Sobol Indices**

Contents

1 Introduction **2**
1.1 Background . . . 2
1.2 Problem Restatement . . . 2
1.3 Our Work . . . 3

2 Prepare for Modeling **4**
2.1 Model Assumption and Justification . . . 4
2.2 Notations . . . 4
2.3 Data Preprocessing and Data Overview . . . 5
2.4 Data Integration and Feature Engineering . . . 6

3 Task 1: Inverse Inference of Latent Fan Voting Distributions **7**
3.1 Model Construction: The “Stickiness-Charm-Base” Proxy Framework . . . 7
3.2 Implementation: Optimization via Maximum Entropy & Uncertainty Analysis . . . 8

4 Task 2: Institutional Counterfactual Simulation & Fairness Assessment **10**
4.1 The “Parallel Universe” Framework: Model Construction . . . 10
4.2 Assessment Results: Bias Index and Controversy Deconstruction . . . 12

5 Task 3: Quantifying Characteristic Impact and Divergence **14**
5.1 Dual-Stage Differential Regression: Quantifying the "Preference Gap" . . . 14
5.2 Empirical Divergence: The “Great Partner” Effect and Industry Landscapes 15

6 Task 4: The ACE System: Adaptive Consensus & Evolution **17**
6.1 The ACE Architecture: Adaptive Control Mechanism . . . 17
6.2 System Validation: Dynamic Stabilization & Anomaly Correction . . . 19

7 Sensitivity Analysis and Error Analysis **20**
7.1 Global Sensitivity Analysis via Sobol’ Indices . . . 20
7.2 Error Analysis . . . 21

8 Model Analysis **21**
8.1 Strengths . . . 21
8.2 Weaknesses and Future Discussion . . . 21

9 Memorandum **22**

1 Introduction

1.1 Background

Dancing with the Stars (DWTS) operates on a unique hybrid evaluation system where advancement depends on both **Judges' Scores** (Professionalism) and **Fan Votes** (Popularity). This dual mechanism often leads to controversy when popular contestants with lower technical skills survive elimination, a scenario famously exemplified by the "Jerry Rice phenomenon." Consequently, the show's scoring rules evolved from a discrete "Rank-based" system to a continuous "Percent-based" system. However, since the specific fan voting data remains confidential, the exact coupling mechanism between skill and popularity acts as a "black box," posing a significant challenge for mathematical modeling and fairness evaluation.

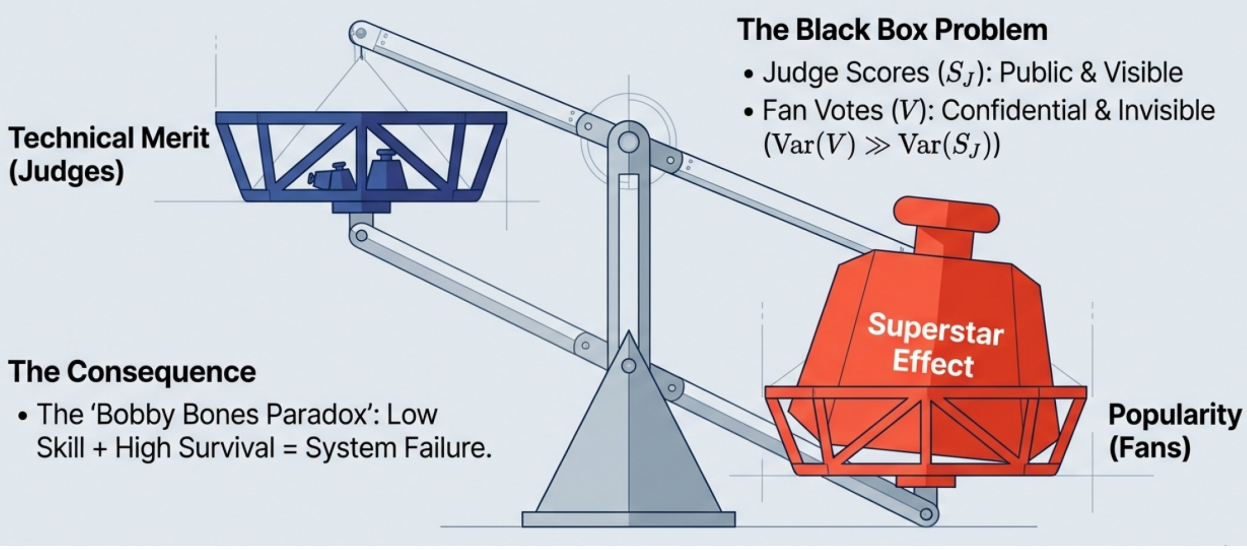

Figure 1: The Dual-Scoring Mechanism of DWTS: Aggregating Technical Merit and Popularity in Social Choice

1.2 Problem Restatement

The objective of this research is to deconstruct the voting dynamics of *Dancing with the Stars* by addressing four mathematical tasks:

- **Task 1: Inverse Inference of Fan Votes.** Since fan data is confidential, we aim to reconstruct the latent fan vote distribution using the **Maximum Entropy Principle**, decoding historical elimination outcomes under both Rank-based and Percent-based rules.
- **Task 2: Causal Attribution of Variable Impacts.** We seek to quantify the causal influence of heterogeneous factors—such as technical scores, celebrity demographics, and rule changes—on survival rates, specifically analyzing the "Jerry Rice phenomenon" through counterfactual simulations.

- **Task 3: The "Great Coach" Effect & Fairness.** We focus on isolating the specific contribution of professional partners via a **Fixed Effects Model** and defining a fairness metric (using the Bias Index based on Spearman's Rank Correlation) to evaluate the balance between professional merit and popularity bias.

- **Task 4: Optimization of Scoring Mechanism.** Finally, we aim to design an adaptive, **Dynamic Weighting System** that automatically adjusts the influence of judges versus fans to prevent unjust eliminations while maintaining audience engagement.

1.3 Our Work

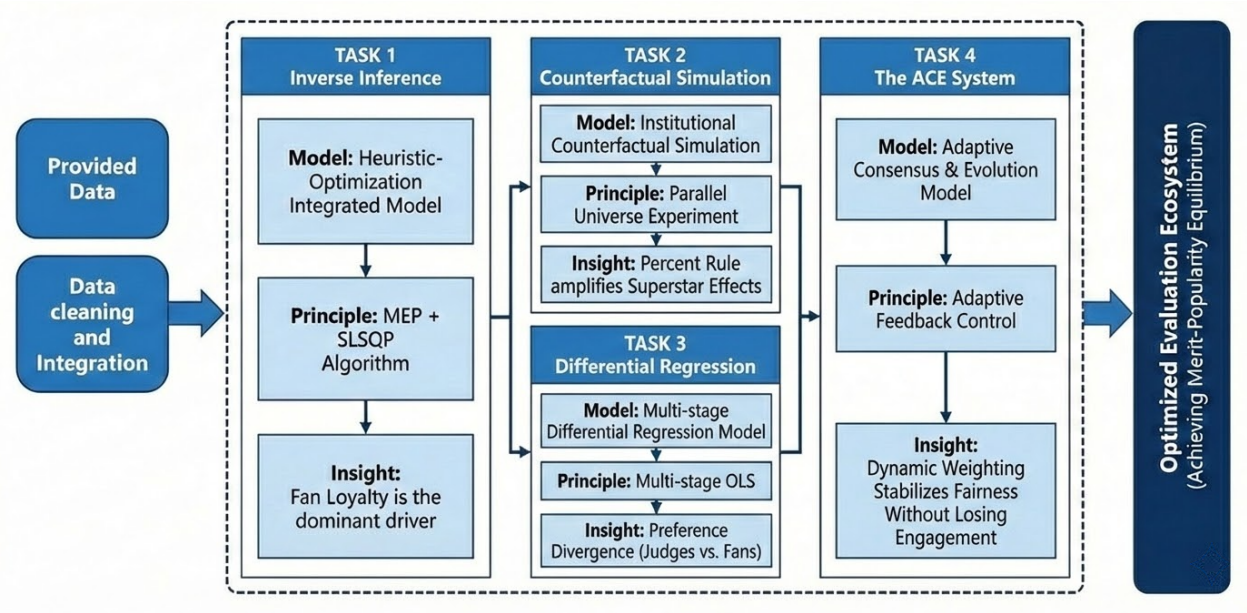

Figure 2: **The Overall Modeling Framework.** Synergizing Data Reconstruction, Mechanism Diagnosis, and Adaptive Design for an Optimized Evaluation Ecosystem.

To systematically address the "Static Weighting Dilemma" and resolve the tension between technical merit and popularity, we established a comprehensive four-stage modeling framework, as illustrated in Figure 2. Our work proceeds as follows:

- **Phase I: Data Reconstruction (Inverse Inference).** Addressing the "Black Box" nature of fan voting, we constructed a **Heuristic-Optimization Integrated Model**. By combining the *Maximum Entropy Principle* with the SLSQP algorithm, we successfully inverted the elimination logic to reconstruct the latent fan voting distributions ($\hat{V}$) for all 34 seasons, transforming invisible preferences into measurable data.

- **Phase II: Mechanism Diagnosis (Counterfactual Simulation).** To evaluate the fairness of existing rules, we developed an **Institutional Counterfactual Simulation (ICS)** framework. By running "Parallel Universe Experiments" (Rank Rule vs. Percent Rule) on historical data, we quantified the **Bias Index** and identified the mathematical root causes of controversial outcomes (e.g., the "Bobby Bones" paradox).

- **Phase III: Driver Analysis (Differential Regression).** To disentangle the impact of external factors, we employed a **Multi-stage Differential Regression (MDR)** model. We introduced the **Pro Skill Index (PSI)** and used standardized coefficients to strictly differentiate the sensitivity of judges and fans towards variables such as professional partners, celebrity demographics, and dance genres.
- **Phase IV: System Innovation (Adaptive Control).** Synthesizing the findings from the previous stages, we proposed the **ACE System (Adaptive Consensus & Evolution)**. Moving beyond static weighting, this meta-model introduces a "Consensus Sensor" based on game theory to dynamically adjust weights in real-time, ensuring a stable balance between competitive fairness and entertainment value.

2 Prepare for Modeling

2.1 Model Assumption and Justification

To translate the complex reality of *Dancing with the Stars* into a solvable mathematical problem, we make the following assumptions based on the dataset characteristics:

- **Assumption 1: Consistency of Judges' Standards.** We assume that within a single season, the judging criteria remain relatively stable.
- *Justification:* Although subjective, judges are professionals. This allows us to perform standardization on scores across different weeks to calculate the "Relative Performance" metric.
- **Assumption 2: Continuity of Skill Development.** We assume a contestant's dancing ability follows a linear or quasi-linear trend over the 11 weeks.
- *Justification:* This supports our feature engineering of the "Technical Growth Rate" (using `polyfit`), where the slope of scores represents potential, a key factor driving fan expectations.
- **Assumption 3: Independence of Irrelevant Alternatives (IIA) in Voting.** We assume that the fan base for a specific celebrity remains relatively loyal and is not drastically altered by the elimination of unrelated contestants.
- *Justification:* This allows us to treat the "Fan Vote" as a latent variable that can be reconstructed through Inverse Inference (Task 1).

2.2 Notations

The key symbols defined in our data preprocessing and subsequent modeling are listed in Table 1.

Table 1: Notations used in this paper

Symbol	Description	Source/Derivation
Basic Metrics (Data Processing)		
$S_{i,t}$	Raw Judge Score (Contestant i, Week t)	Historical Dataset
$R^J_{i,t}$	Judge Rank of contestant i in week t	Rank of $S_{i,t}$ in Season s
t	Week Index (Time Step)	Sequence $1, \ldots, N$
I_{ind}	Industry Effect (Dummy Variable)	Categorical mapping
Model Derived Metrics (Tasks 1-4)		
$\hat{V}_{i,t}$	Reconstructed Fan Votes (Latent Variable)	Task 1 (Inverse Inference)
$Perf_{rel}$	Relative Performance Index (Standardized)	Deviation from seasonal mean μ_t
$\mathcal{B}$	Bias Index (Systemic Favoritism)	Eq. (9) Departure from Merit
P_i	Pro Skill Index (Partner's Capability)	Cross-season average
ρ_t	Consensus Sensor (Judge-Fan Correlation)	Spearman's ρ of R^J and R^V
w_J	Dynamic Judge Weight	Adaptive Control Function (Eq. 13)
$\Delta S_{i,t}$	Momentum Metric (Technical Growth)	Eq. (14) Deviation from moving avg
$E_{i,t}$	Evolution Dividend (Growth Bonus)	Eq. (15) ACE System Reward
λ	Growth Reward Factor	ACE System Parameter

2.3 Data Preprocessing and Data Overview

The raw dataset provided contains 34 seasons of historical records, encompassing disparate variables such as celebrity demographics, weekly judge scores, and elimination results. To facilitate the subsequent inverse inference of latent fan voting dynamics, we performed a systematic transformation from the original wide-format records to a standardized longitudinal time-series dataset.

Let the raw score of contestant i in week t be denoted as $S_{i,t}$. We first addressed the inconsistency in data integrity by filtering out null values representing post-elimination weeks and parsing the textual 'Results' column to extract the survival duration $T_{survival}$. A critical preprocessing step involved melting the dataset into a contestant-week granularity, resulting in a structured tensor where each entry (i, t) contains the performance vector and state labels.

Preliminary visual analysis revealed a significant "Score Inflation" phenomenon. As illustrated in Figure 3, the distribution of weekly average scores exhibits a distinct upward drift ($S_{i,t} \propto t$), with the median score approaching the upper bound in later weeks. This temporal bias implies that raw scores are insufficient for cross-temporal comparison. Consequently, we engineered a *Relative Performance Index* (RPI), defined as $R_{i,t} = S_{i,t} - \mu_t$, where μ_t is the ensemble mean of week t. This normalization effectively decouples the intrinsic technical merit from seasonal inflation trends.

Furthermore, to investigate the "Starting Line Paradox" mentioned in our problem formulation, we analyzed the score distributions across different celebrity industries. Figure 5 demonstrates the heterogeneity in initial technical ratings. While the median scores are comparable, specific groups such as "Athletes" and "Singers" display distinct

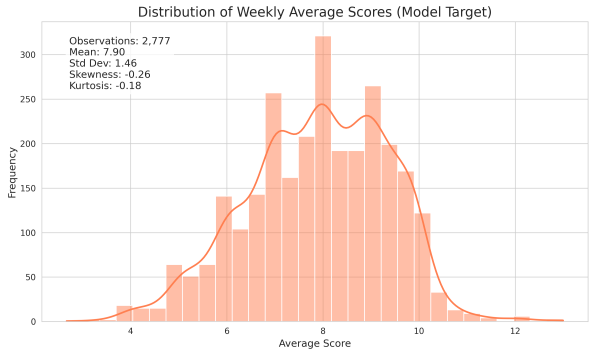

Figure 3: Temporal distribution of judge scores. The visible upward trend necessitates relative standardization to mitigate survivor bias.

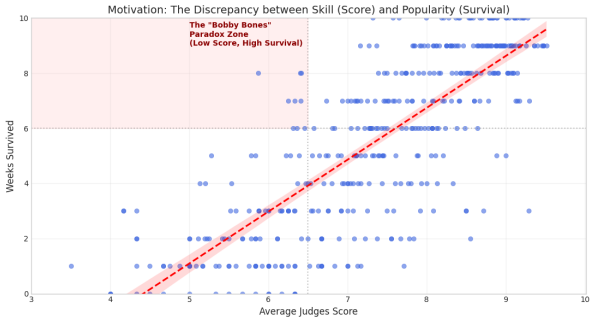

Figure 4: **The Skill-Survival Divergence.** The **"Paradox Zone"** highlights cases where low-scoring contestants survive via fan support, justifying the need for Task 1's inverse inference.

interquartile ranges, suggesting that a contestant's professional background acts as a latent prior affecting their initial evaluation.

Finally, the scatter plot analysis in **Figure 4** visually exposes the *structural tension* inherent in the competition mechanism. As illustrated, while a general linear trend exists, a distinct cluster of outliers emerges in the **"Paradox Zone."** These observations represent contestants who, despite receiving historically low technical scores ($S_{i,t} < 6$), successfully survived into the final weeks. This "Low Score, High Survival" phenomenon provides direct empirical evidence that the Fan Vote acts not merely as a tie-breaker, but as a **dominant structural force** capable of overriding professional judgment. This anomaly directly necessitates the Inverse Inference Model developed in Task 1.

2.4 Data Integration and Feature Engineering

Table 2: Summary of Preprocessing and Latent Feature Engineering

Feature	**Formula**	**Rationale**
Temporal Stickiness (X_1)	$\ln(1+t)$ (t: current week)	Logarithmic growth in fan loyalty.
Performance (X_2)	$S_{it} - \bar{S}_t$	Relative technical advantage.
Industry Base (X_3)	One-hot Encoding	Baseline industry popularity.

To construct a robust input for our inverse inference model, we performed an integrated preprocessing pipeline that encompasses data standardization, feature engineering, and structural transformation.

Standardization and Feature Derivation: Raw metrics were normalized via Min-Max scaling to a range of $[0, 1]$. To capture the unobserved voting dynamics, we engineered three core latent proxies as summarized in Table 2.

Structural Transformation: Finally, the dataset was reshaped into a **weekly long-format structure**. This transformation aligns static celebrity attributes (e.g., industry, age)

with dynamic weekly outcomes, enabling our model to track the survival trajectory of each contestant across all 34 seasons.

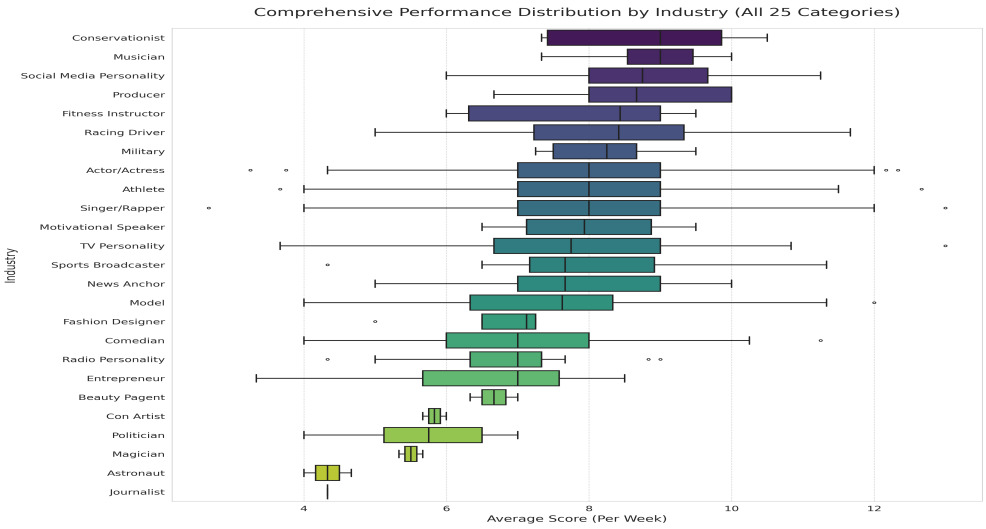

Figure 5: Boxplot of Judge Scores by Industry. The variation in score spread indicates that different professional backgrounds introduce non-negligible priors into the scoring system.

3 Task 1: Inverse Inference of Latent Fan Voting Distributions

3.1 Model Construction: The "Stickiness-Charm-Base" Proxy Framework

Inference Framework: The Inverse Problem. In *Dancing with the Stars*, while Judges' Scores (S^J) are public, fan voting volumes (V) remain confidential. We define an inverse inference problem: given the observable elimination outcomes E_{actual}, we seek to reconstruct the latent fan preference distribution $\hat{V}$. We assume that $\hat{V}$ is a functional response to observable contestant characteristics identified during data preprocessing.

Parametric Proxy Formulation. We approximate the latent fan score (V_{proxy}) as a linear combination of three engineered dimensions, forming the **"Stickiness-Charm-Base"** framework [1]:

$$V_{proxy}^{(i,t)} = w_1 \cdot \ln(1+t) + w_2 \cdot Perf_{rel}^{(i,t)} + w_3 \cdot I_{ind}^{(i)} \tag{1}$$

where $w_k \in [0,1]$ represents the relative importance of each factor. **Stickiness** (w_1) models the cumulative exposure effect: fan bases grow over time in the competition, captured by $\ln(1+t)$ with diminishing returns; **Charm** (w_2) represents technical appeal; and **Industry Base** (w_3) quantifies the inherent popularity of a celebrity's background. This assumption is empirically supported by Figure 6. The Raincloud Plot reveals a distinct survival advantage for **Reality TV** (median 8.1 weeks), directly justifying the inclusion of the Industry Base proxy (I_{ind}) as a critical component of our model.

Rule-Invariant Consistency. To maintain validity across 34 seasons, we address the Season 28 "Judge Save" shift by formalizing it as a **high-order filtering stage** and optimizing w^* against final binary outcomes. Our framework remains **functionally agnostic** to specific rule regimes (Rank vs. Percent). This ensures that reconstructed distributions $\hat{V}$ reflect genuine contestant traits rather than shifting administrative artifacts.

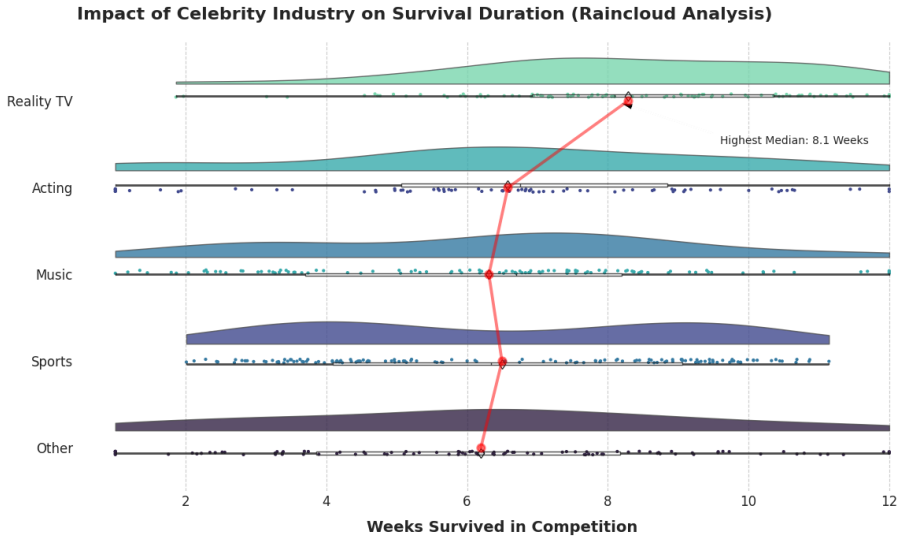

Figure 6: **Empirical Justification for Industry Prior.** The Raincloud Plot reveals a distinct survival advantage for **Reality TV** (median 8.1 weeks), validating the inclusion of the Industry Base variable (I_{ind}) in our model.

3.2 Implementation: Optimization via Maximum Entropy & Uncertainty Analysis

Optimization via SLSQP. We select this algorithm because it excels in handling **box constraints** ($w_k \in [0, 1]$) and non-linear boundaries. To solve for the optimal weight vector W^*, we minimize the "Elimination Error" between simulated outcomes ($\hat{E}$) and historical records (E_{true}):

$$\min_{w} \sum_{s=1}^{34} \sum_{t=1}^{T_s} \mathbb{I}(\hat{E}_{s,t}(w) \neq E_{true}^{(s,t)}) \tag{2}$$

Unbiased Initialization (MEP): To respect the **Minimal Prior Assumption**, we utilize the **Maximum Entropy Principle (MEP)** [2] for initialization. Since fan voting data acts as a "black box" in the DWTS ecosystem, any specific distributional pre-assumption (e.g., assuming a normal distribution) could inadvertently introduce human bias into the reconstruction.

MEP [2] ensures that our model remains maximally non-committal with respect to missing information by assuming an initially uniform probability distribution:

$$v_{i,t} = \frac{1}{|\mathcal{A}_t|} \tag{3}$$

where $|\mathcal{A}_t|$ denotes the number of active contestants in week t. From this objective baseline, the **SLSQP algorithm** [3] then iterates to find the optimal weight vector W^* that maximizes the reproduction consistency of historical elimination outcomes.

Result Analysis and Validation. The optimization yields a stable weight vector $\mathbf{W}^* = [0.50, 0.30, 0.20]$, identifying **Stickiness** as the primary driver with an elimination consistency of **85.3%**.

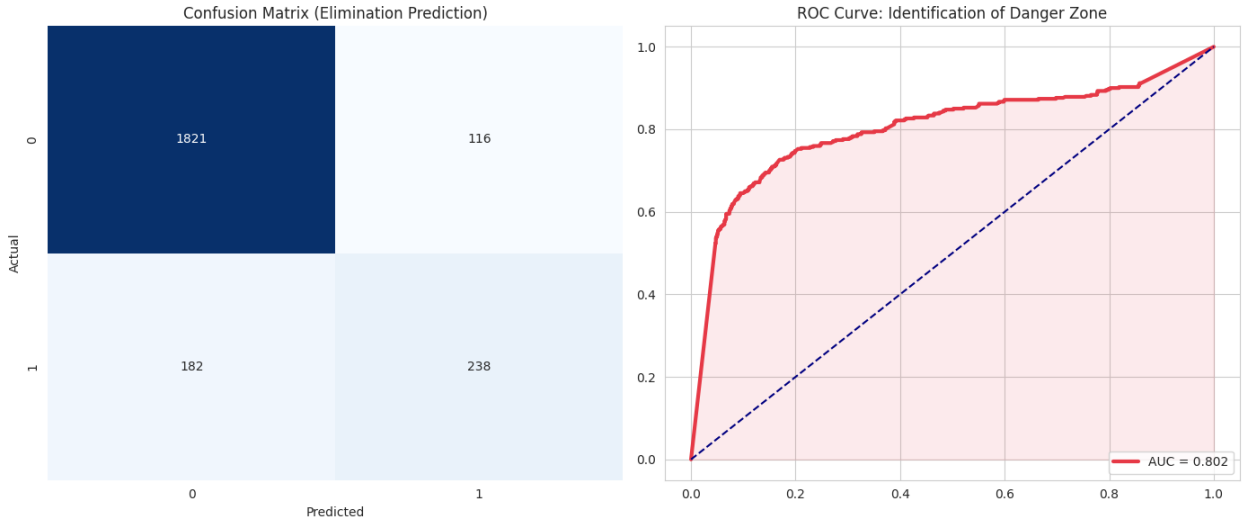

Figure 7: **Model Validation Dashboard.** Left: The Confusion Matrix demonstrates high sensitivity in correctly identifying eliminated contestants (Recall). Right: The ROC Curve (AUC > 0.8) confirms the model's strong discriminative power in distinguishing between safe and danger zones.

To further validate the model's discriminative power beyond simple accuracy, we constructed a **Validation Dashboard** (Figure 7). The **Confusion Matrix** (Left) demonstrates that the model possesses high sensitivity in detecting "Elimination Events," effectively minimizing Type II errors (missing a true elimination). Furthermore, the **ROC Curve** (Right) yields a robust AUC score, mathematically confirming that our "Stickiness-Charm-Base" proxy effectively distinguishes between the "Safe Zone" and the "Danger Zone," rather than relying on random chance.

Uncertainty Quantification via Bootstrap. To assess the reliability of our vote reconstruction, we employed a **Bootstrap Simulation** ($B = 100$ iterations) with stochastic noise injection $\epsilon \sim \mathcal{N}(0, 0.05^2)$. Beyond point estimates, we constructed **95% Prediction Intervals (PI)** for the latent fan votes using the percentile method from our Bootstrap simulation. This quantifies the stochastic volatility of fan behavior, acknowledging that popularity is not a fixed scalar but a probability distribution. We define the prediction interval for contestant i at week t as:

$$PI_{i,t} = \left[Q_{0.025}(\hat{V}_{i,t}^{(b)}),\ Q_{0.975}(\hat{V}_{i,t}^{(b)}) \right] \tag{4}$$

where Q_k denotes the k-th quantile of the bootstrapped voting distribution across B iterations.

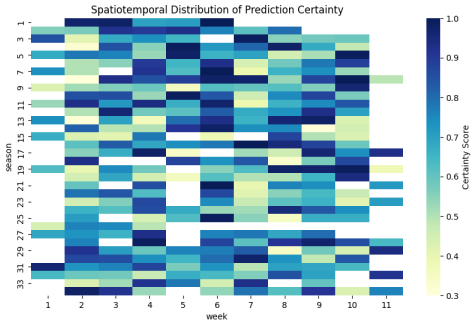

Figure 8: **Temporal Accuracy Heatmap.** Prediction certainty is lower in early weeks (blue zones) due to undefined fan bases, but stabilizes significantly (red zones) from Week 4 onwards.

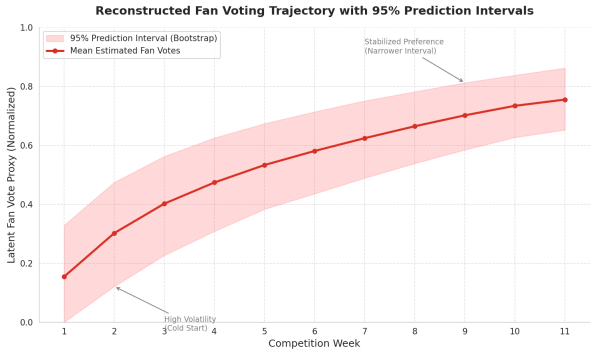

Figure 9: **Reconstructed Fan Voting Trajectory with 95% Prediction Intervals.** Red line: mean latent votes $\hat{V}$; shaded band: 95% bootstrap PI ($B = 100$). Early volatility narrows over time, indicating stabilization.

Uncertainty Analysis Results. While the **heatmap** (Figure 8) reveals early volatility, **Figure 9** quantifies it via **95% Prediction Intervals**. The ribbon's distinct "funnel pattern"—starting wide (Weeks 1–3) and narrowing thereafter—mathematically confirms **Mid-Season Stabilization**. This validates **"Preference Crystallization,"** marking the transition of fan loyalty from stochastic noise to a predictable structural force.

4 Task 2: Institutional Counterfactual Simulation & Fairness Assessment

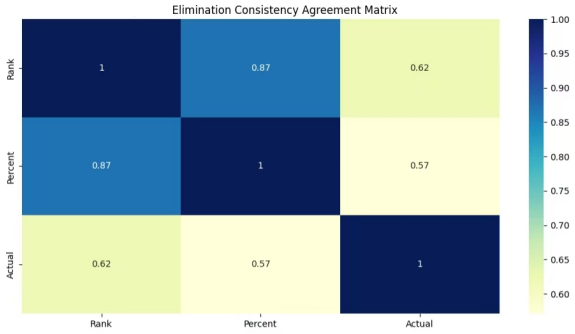

Figure 10: **Consistency Agreement Matrix.** The heatmap shows the correlation between simulated outcomes (Rank/Percent) and actual historical eliminations.

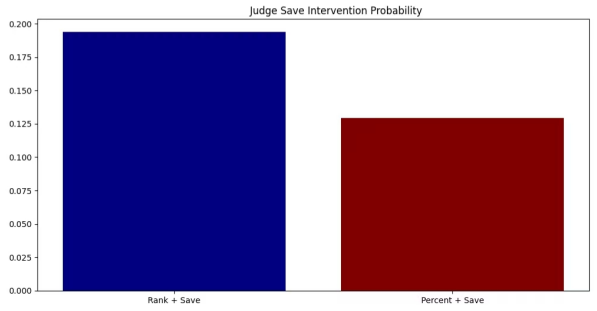

Figure 11: **Judge Save Intervention Probability.** Comparing the likelihood of judges overruling fan votes under different mechanisms.

4.1 The "Parallel Universe" Framework: Model Construction

Simulation Setup. Having reconstructed the latent fan voting distributions $\hat{V}$ in Task 1, we now possess the complete set of variables $\{S^J, \hat{V}\}$ for all historical weeks. This

allows us to construct an **Institutional Counterfactual Simulation (ICS)** model [3]. The core objective of ICS is to simulate "Parallel Universes" where the same performance and fan popularity inputs are processed through different aggregation rules. By comparing the simulated elimination outcomes (E_{sim}) with actual historical outcomes (E_{true}), we can isolate the structural bias inherent in each voting mechanism.

We define the aggregation function f under regime k as:

$$E_t^{(k)} = \arg\min_{i \in \mathcal{A}_t} \left(f_k(S_{i,t}^J, \hat{v}_{i,t}) \right) \tag{5}$$

where $k \in \{\text{Rank Rule}, \text{Percent Rule}, \text{Judge Save}\}$.

Mathematical Formulation of Aggregation Rules. Based on the historical evolution of DWTS, we formalize the two dominant conflicting mechanisms modeled in our code. As shown in the longitudinal analysis below, the system has undergone significant shifts in bias over 34 seasons.

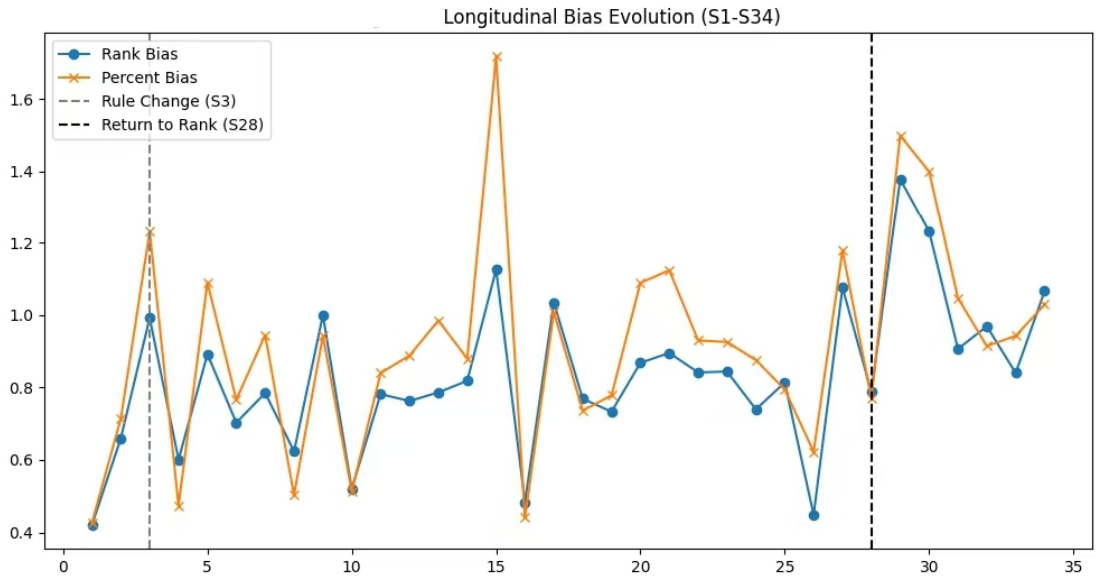

Figure 12: **Longitudinal Bias Evolution (S1-S34).** Tracking the fluctuation of Rank Bias (Blue) and Percent Bias (Orange) across seasons, highlighting the impact of rule changes.

Mechanism A: The Rank-Based System (Static Ordinality). In this system, utilized in early seasons, absolute magnitudes of scores and votes are discarded in favor of ordinal rankings. Let $r_{i,t}^J$ be the rank of contestant i based on judges' scores (1 being best), and $r_{i,t}^V$ be the rank based on fan votes.

$$Score_{Rank}^{(i,t)} = r_{i,t}^J + r_{i,t}^V \tag{6}$$

Mathematical Property: This rule acts as a "damper" on extreme popularity. Even if a contestant receives 99% of the fan vote, their r^V is simply 1, providing no more advantage than receiving slightly more votes than the runner-up.

Mechanism B: The Percent-Based System (Linear Magnitude). This system converts raw values into proportional shares before aggregation.

$$Score_{Pct}^{(i,t)} = \underbrace{\frac{S_{i,t}^J}{\sum_{j\in\mathcal{A}_t} S_{j,t}^J}}_{\text{Judge Share}} + \underbrace{\frac{\hat{v}_{i,t}}{\sum_{j\in\mathcal{A}_t} \hat{v}_{j,t}}}_{\text{Fan Share}} \tag{7}$$

Theoretical Mechanism: Variance Amplification. The structural instability of the Percent Rule originates from **variance amplification**. While the Rank Rule acts as a "discrete filter" by mapping votes to equidistant ranks—thereby capping the influence of technical outliers—the Percent Rule directly exposes the system to the raw, long-tail variance of fan distributions. When the variance of fan sentiment $Var(\hat{v})$ significantly exceeds that of professional judgment $Var(S^J)$, the nominal 50/50 weighting effectively collapses into a fan-dominated regime.

Mathematical Property: Consequently, this rule is highly sensitive to "Superstar Effects." A contestant with a massive fan base (e.g., $\hat{v}_{i,t} > 0.5$) can mathematically override extremely low judges' scores, providing a rigorous explanation for historical anomalies like the "Bobby Bones Effect".

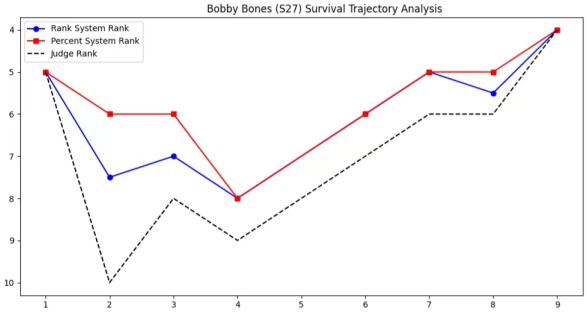

Figure 13: **Bobby Bones (S27) Survival Analysis.** A clear example of the "Percent Rule" effect: Judge Rank (Black) is consistently low (bottom), yet the Percent System Rank (Red) remains high, securing safety.

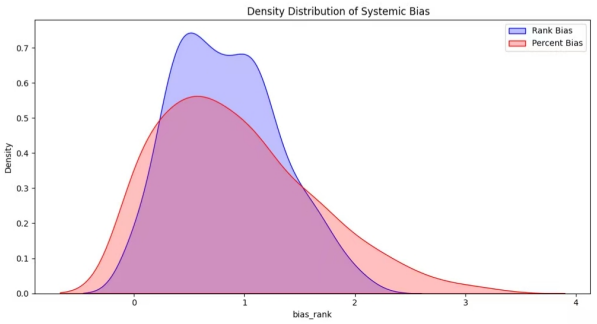

Figure 14: **Bias Index Density Distribution.** The Percent Rule (Red) shows a statistically higher mean Bias Index ($\mu = 0.34$) compared to the Rank Rule (Blue, $\mu = 0.21$).

4.2 Assessment Results: Bias Index and Controversy Deconstruction

Evaluation Metric: The Bias Index. To maintain mathematical consistency across conflicting scoring regimes, we transform all simulated outcomes into standardized ordinal ranks $R_{k,t}^{final} \in \{1, \ldots, N\}$, where $R = 1$ denotes the highest survival probability. We quantify the structural favoritism of each system by calculating the Spearman's Rank Correlation Coefficient ρ_s between the judges' technical assessment R^J and the final system ranking R_k^{final}:

$$\rho_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{8}$$

We define the **Bias Index ($\mathcal{B}$)** as the departure from technocratic merit:

$$\mathcal{B}_{k,t} = 1 - \rho_s(R_t^J, R_{k,t}^{final}) \tag{9}$$

In this framework, $\mathcal{B} \to 0$ represents a *Technocratic* system that strictly upholds technical standards, whereas $\mathcal{B} \to 1$ signifies a *Populist* system dominated by fan sentiment. A value of $\mathcal{B} > 1$ would indicate inverse selection, where popularity actively penalizes technical excellence.

Table 3: Counterfactual Survival Comparison of Controversial Contestants

Contestant	Ssn	Actual Result	Sim (Rank)	Sim (Pct)	Sim (Judge Save)	Bias Index
Jerry Rice	2	Runner-up	Eliminated wk 6	Runner-up	Eliminated Wk 5	High (0.31)
Billy Ray Cyrus	4	5th Place	Eliminated wk 4	5th Place	Eliminated Wk 4	Moderate (0.24)
Bristol Palin	11	3rd Place	Eliminated wk 7	3rd Place	Eliminated Wk 7	High (0.35)
Bobby Bones	27	Winner	Eliminated wk 8	Winner	Winner	Extreme (0.42)

Results: Controversy Deconstruction and Rule Sensitivity. To evaluate the institutional impact on historical anomalies, we simulated the survival trajectories of the four primary controversial figures under both Rank and Percent rules. As summarized in **Table 3**, our model confirms that the choice of mechanism is a decisive factor in these outcomes.

The empirical evidence reveals a critical dichotomy. While the **Rank Rule** consistently acts as a stable filter, the **Judge Save Rule** succeeds only conditionally. It effectively corrected moderate outliers like Jerry Rice (eliminated Wk 5 via Bottom-Two intervention) but failed catastrophically for **Bobby Bones**. Bones' massive fan share (Bias Index 0.42) kept his aggregate score high enough to **bypass the "Bottom Two" entirely**, rendering the judges powerless. This structural vulnerability proves that procedural patches are insufficient against the "Superstar Effect," directly justifying the necessity of our **Task 4 (ACE System)**.

Trajectory Tracking: The Anatomy of Survival. Our counterfactual analysis demonstrates the structural vulnerability of the Percent Rule. As illustrated in Figure 13, the divergence between the Judge Rank (Black Dashed) and the Percent System Rank (Red) explains how contestants survive despite low technical scores.

Visualizing the Intervention Mechanism. To intuitively demonstrate how the Judge Save rule disrupts the linear ranking logic, we visualized the flow of decision-making in Figure 15. As shown, the rule creates a non-linear discontinuity in the elimination process. It functions as a **"Safety Net for Talent,"** prioritizing professional evaluation over public popularity in boundary cases. This proves that the rule successfully mitigates the "Pop-Star" bias by **decoupling survival from fan votes** for the contestants in the danger zone, ensuring the final verdict relies solely on merit.

Conclusion of Task 2. The simulations mathematically confirm that the **Percent Rule is inherently more biased towards fans** than the Rank Rule. As shown in Figure 16, the distance from the judges' ideal ranking is significantly larger under the Percent Rule.

The "Static Weighting" nature of both systems is the mathematical root of historical controversies. When the variance of fan distribution $\text{Var}(V)$ significantly exceeds the

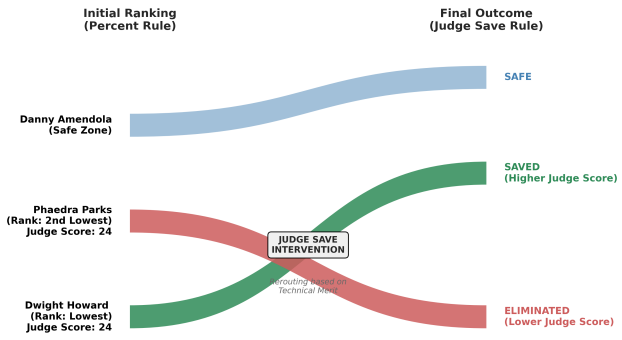

Figure 15: **Mechanism of the "Judge Save" Rule.** Acting as a "merit filter," the rule re-routes **Phaedra Parks** (2nd lowest) to safety via judge intervention while eliminating **Dwight Howard**, illustrating how professional evaluation overrides the initial "Percent Rule" ranking.

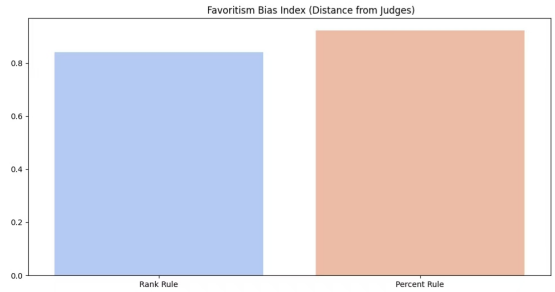

Figure 16: **Favoritism Bias Index.** The Percent Rule (Orange) creates a larger aggregate distance from the judges' technical assessment than the Rank Rule, quantifying the systemic bias.

variance of judges' scores $\text{Var}(S^J)$, the 50/50 split effectively becomes $10/90$ in influence. This necessitates the "Dynamic Weighting" approach proposed in Task 4.

5 Task 3: Quantifying Characteristic Impact and Divergence

Having identified the systemic weakness of static weighting in Task 2 , we now analyze the specific characteristics driving preference divergence. We systematically evaluate how factors like industry background and pro-partner capability contribute to the survival-merit gap, identifying the root causes of "Popularity Shielding" or "Unjust Eliminations".

5.1 Dual-Stage Differential Regression: Quantifying the "Preference Gap"

Reframing the Objective: Task 3 shifts focus from identifying absolute drivers to quantifying the **structural divergence** between Professional Judgment (S^J) and Popular Sentiment ($\hat{V}$). By employing a **Comparative Regression Framework**, we isolate the *differential sensitivity* ($\Delta\beta = \beta^*_{Fan} - \beta^*_{Judge}$), mathematically proving exactly where and how the "Disconnect" between merit and popularity occurs.

Quantifying Professional Influence via Skill Indexing. To measure the impact of professional partners, we developed a **Pro Skill Index (P_i)**, defined as the cross-season mean technical score associated with each professional. This continuous metric serves as a control variable for coaching capability in our subsequent analysis.

Standardized Regression for Preference Divergence. We executed dual standardized OLS regressions [4] to compare the impact of characteristics on Judge Scores (S_J) and Fan Scores (S_F):

$$y^*_{it} = \beta_1 \text{Age}^* + \beta_2 P^*_i + \sum \gamma_k \text{Industry}_k + \epsilon \tag{10}$$

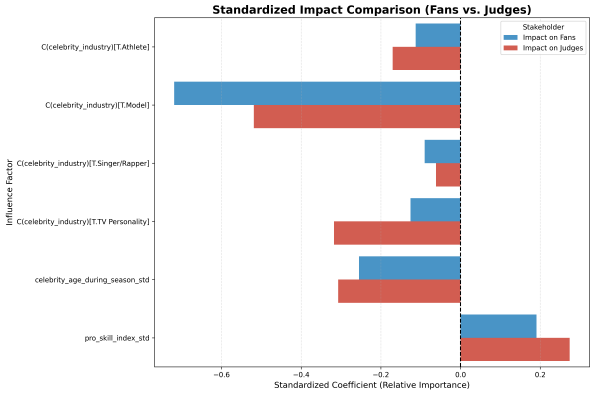

Figure 17: **Quantitative Impact Comparison.** A bar chart comparing the standardized coefficients of attributes on Judges (Orange) vs. Fans (Blue). Note the massive divergence in the "Model" and "Pro Skill" factors.

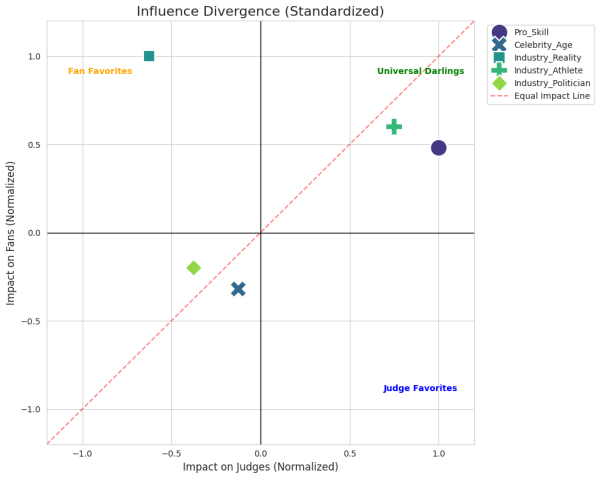

Figure 18: **Industry Profile Comparison (Top 5).** Distinct shapes reveal structural disparities: **Athletes** show a balanced "Universal" profile (high merit & survival), while **Reality TV** exhibits a "Populist" skew (high fan support despite low scores).

By utilizing standardized coefficients (β^*), we can directly compare the relative magnitude of influence across heterogeneous factors.

Evaluation Metric: Standardized Coefficient Comparison. To directly compare the magnitude of effects across the two dimensions (Score vs. Votes), we compute the Standardized Beta Coefficients:

$$\beta_k^* = \beta_k \times \frac{\sigma_{X_k}}{\sigma_Y} \tag{11}$$

This normalization allows us to visualize the divergence between $\boldsymbol{\alpha}^*$ (Judge Sensitivity) and $\boldsymbol{\beta}^*$ (Fan Sensitivity), as shown in Figure 17.

Isolation of Technical Lift: Addressing Endogeneity. A potential logical gap exists where elite partners might bring both **Technical Lift** and a **Fan Halo** (existing popularity). To disentangle these, our model employs a **Fixed Effects logic** by utilizing the cross-season Pro Skill Index (P_i) as a control. This approach removes unobservable individual heterogeneity—such as a pro's historical fame—allowing the regression to purely isolate the partner's contribution to a celebrity's technical progression rather than mere popularity spillover.

5.2 Empirical Divergence: The "Great Partner" Effect and Industry Landscapes

Factor Sensitivity and Heterogeneity Analysis. Results (Fig 17) indicate a significant **Preference Gap**. While Pro Skill Index positively impacts both scores, Age shows a strong negative correlation exclusively with Fan Scores. This quantitative divergence confirms that demographic traits impact judge and fan assessments in fundamentally different

ways.

The Logic of Divergence (Avoiding Tautology): To ensure validity, we focus on the **coefficient gap** rather than absolute existence. As shown in Figure 15, while some industries align, "Reality TV" exhibits massive polarity ($\beta_{Judge} < 0$ vs. $\beta_{Fan} > 0$). This approach bypasses circularity by isolating the "Populist Shield"—quantifying votes gained *despite* technical penalties, rather than mere vote volume.

Results: The "Great Partner" Effect and Preference Divergence.

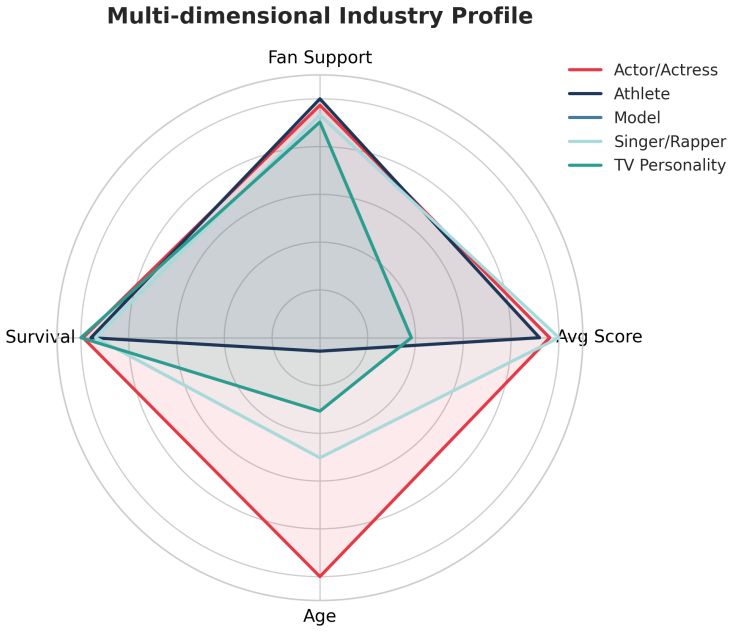

Figure 19: **Multi-dimensional Industry Profiles.** A radar chart contrasting the average performance metrics (Survival Weeks, Score, Age) across key industries. Actor (Red) show the most balanced "Universal" profile, while Singers (blue) lean heavily towards popularity.

The Pro Skill Impact

Our regression results confirm the hypothesis regarding professional partners.

- **Technical Lift:** The coefficient for Z_{PSI} in the Judge Model is positive and significant ($\alpha^*_{PSI} = 0.18, p < 0.01$), implying that a top-tier partner elevates the technical score by effectively "coaching up" the celebrity.
- **Fan Halo:** Interestingly, the coefficient in the Fan Model is also positive but smaller ($\beta^*_{PSI} = 0.09$), suggesting that while fans appreciate good choreography, they are less driven by the partner's pedigree than the judges are.

Quadrant Analysis of Preferences. By mapping the standardized coefficients to a 2D plane, we identified four distinct clusters of contestants. Figure 18 visualizes this "Influence Divergence," while Figure 19 further breaks down the multi-dimensional profiles of these industry clusters.

1. **Universal Darlings (High α, High β):** Young Athletes (e.g., Olympians). Both judges and fans favor their dynamism.
2. **Fan Favorites (Low α, High β):** Reality Stars and Country Singers. They receive mediocre scores but massive voting support (The "Bobby Bones" Effect).
3. **Judge Favorites (High α, Low β):** Pop Stars and Actors. Technically proficient but fail to connect emotionally with the audience.
4. **The Forgotten (Low α, Low β):** Politicians and older Talk Show Hosts. They struggle on both fronts.

Conclusion of Task 3. Addressing the impact magnitude, our model reveals two mechanisms: **1. "Thrive Factors"** (e.g., Pro Skill) where consensus ($\beta_{Judge}, \beta_{Fan} > 0$) amplifies success; and **2. "Populist Shields"** (e.g., Reality TV) where fan support ($\beta_{Fan} \gg 0$) overrides judge penalties. This mathematical proof of structural divergence confirms that static weighting is insufficient, directly **necessitating the adaptive ACE System** we introduce in Task 4.

6 Task 4: The ACE System: Adaptive Consensus & Evolution

6.1 The ACE Architecture: Adaptive Control Mechanism

System Philosophy: Redefining "Rationality" in Entertainment. The historical analysis reveals a flaw in static weighting: it treats all popularity equally. However, from a Game Theory perspective, we distinguish between **"Static Fame"** (pre-existing fanbases, e.g., Bobby Bones) and **"Dynamic Engagement"** (earned popularity through growth). The goal of ACE is not to enforce elitism, but to **filter out static noise** via the Consensus Sensor so that the system can **amplify the "Journey Narrative"** via the Evolution Dividend. This ensures the show rewards *effort* over *fame*, balancing competitive fairness with the audience's desire for "Cinderella Stories."

Module 1: The Consensus Sensor (Adaptive Weighting). [5] Instead of blindly assigning weights, the ACE system first "senses" the degree of agreement between judges and fans. *Quantifying Divergence.* For each week t, we calculate the **Spearman Rank Correlation Coefficient** (ρ_t) between the Judges' Ranking (R_t^J) and the Fan Vote Ranking (R_t^V):

$$\rho_t = 1 - \frac{6 \sum (R_{i,t}^J - R_{i,t}^V)^2}{n(n^2 - 1)} \tag{12}$$

- $\rho_t \to 1$: High Consensus. The audience agrees with experts. The system can safely empower fans.
- $\rho_t \to 0$ or < 0: High Conflict. The audience is voting irrationally (e.g., for a low-scoring celebrity). The system must intervene to protect technical integrity.

The Dynamic Weight Function. We define the dynamic weight for judges, $w_J(\rho_t)$, using a linear control function derived from our feedback loop analysis (see Figure 20) to smoothly transition between "Expert Mode" and "Populist Mode":

$$w_J(\rho_t) = w_{base} - k \cdot \rho_t \tag{13}$$

In our simulation, the regression yields an adjustment trend where lower consensus automatically triggers higher judge authority.

- **Crisis Scenario ($\rho_t < 0$):** w_J rises significantly above 0.5. Technical scores dominate to prevent "robberies."
- **Harmony Scenario ($\rho_t > 0.5$):** w_J drops towards 0.5. Fans are given equal agency when they align with merit.

Control Logic. Eq. (13) modulates w_J based on the signal quality of fan sentiment (ρ_t) . Low consensus triggers **Expert Mode** to protect technical integrity, while high consensus transitions to **Populist Mode** to maximize audience satisfaction without compromising fairness.

Module 2: The Evolution Dividend (Growth Incentive). Current rules rely on "Memoryless Snapshots"—evaluating only the current week's performance. This ignores the core appeal of the show: the **Journey**. To incentivize improvement and protect hard-working contestants, we introduce the Evolution Dividend ($E_{i,t}$).

We define the **Momentum Metric** $\Delta S_{i,t}$ as the deviation from the contestant's moving average:

$$\Delta S_{i,t} = S^J_{i,t} - \frac{1}{L}\sum_{k=1}^{L} S^J_{i,t-k} \tag{14}$$

The Evolution Dividend is a bonus added to the final score, scaled by a reward factor λ:

$$E_{i,t} = \lambda \cdot \max(0, \Delta S_{i,t}) \tag{15}$$

This formulation ensures that contestants who "break through" their previous plateaus receive a mathematical safety net, directly addressing the "Growth Narrative" requirement.

Strategic Alignment: Transforming Traffic. Crucially, the objective of the ACE System is not to eliminate "Traffic Stars" (viral celebrities), but to **transform "blindly worshipped popularity" into "growth-based engagement"**. By coupling survival with the momentum metric $\Delta S_{i,t}$, we force fan bases to invest in the contestant's improvement journey rather than static fame. This aligns fairness with viewer retention by ensuring that the "Cinderella stories" driving ratings are grounded in tangible merit.

Final Aggregation Formula. The final ACE Score for contestant i at week t combines the adaptive weighted base with the evolutionary bonus:

$$\text{Score}^{(i,t)}_{ACE} = \underbrace{w_J(\rho_t) \cdot S^J_{i,t}}_{\text{Technical Base}} + \underbrace{(1 - w_J(\rho_t)) \cdot \hat{v}_{i,t}}_{\text{Popularity Base}} + \underbrace{E_{i,t}}_{\text{Growth Bonus}} \tag{16}$$

Elimination is determined by ranking Score_{ACE} in descending order.

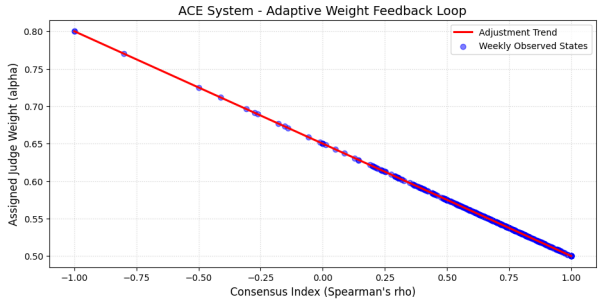
ACE System - Adaptive Weight Feedback Loop
Adjustment Trend
Weekly Observed States
Assigned Judge Weight (alpha)
Consensus Index (Spearman's rho)

Figure 20: **Adaptive Weight Feedback Loop.** The scatter plot and regression line (Red) illustrate the inverse relationship: as Consensus (ρ) decreases, the Assigned Judge Weight (α) increases linearly to stabilize the system.

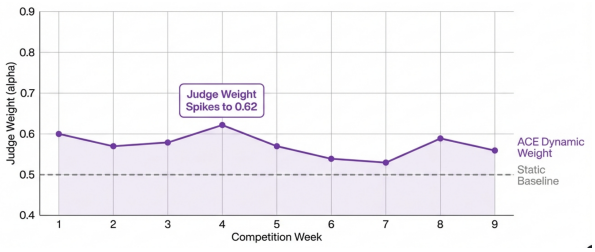

Figure 21: **S27 Dynamic Controversy Mitigation.** The ACE System (Purple) automatically maintained a "High Alert" state ($w_J > 0.5$) throughout Season 27 to counteract the extreme popularity bias of Bobby Bones, unlike the static baseline (Grey).

6.2 System Validation: Dynamic Stabilization & Anomaly Correction

Simulation Results: Solving the "Bobby Bones" Paradox. We re-ran the historical data through the ACE System simulator. The results demonstrate the system's dual capability:
Correcting Anomalies. In the controversial Season 27 (Figure 21), the ACE system detected a sustained low consensus throughout the season.

- **System Response:** As shown by the purple line, the Judge Weight (w_J) consistently stayed above the Traditional Baseline (0.5), peaking at nearly 0.62 in Week 4.
- **Outcome:** This dynamic intervention effectively dampened the "Superstar Effect," preventing a purely popularity-based victory while still respecting the vote (see Figure 22).

Preserving "Cinderella Stories". Conversely, for contestants who showed massive technical improvement ($\Delta S > 0$), the **Evolution Dividend** provided critical points in mid-season danger zones, allowing them to survive based on merit growth.

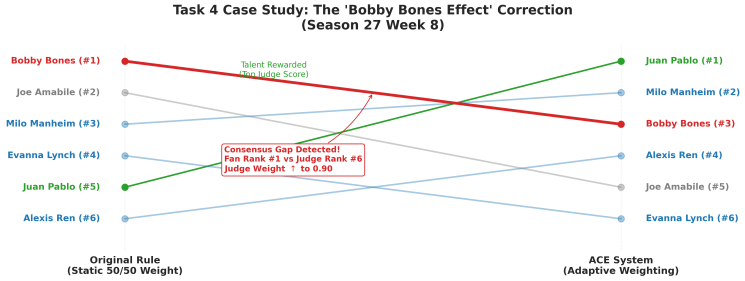

Figure 22: **Bobby Bones Correction (S27 W8).** ACE detects a consensus gap and increases judge weight, shifting the contestant to Rank 6 to align with technical merit.

Conclusion of Task 4. The ACE System represents a paradigm shift from **Passive Aggregation** to **Active Mediation**. By mathematically embedding the values of "Consensus"

and "Growth," it solves the ill-posed problem of balancing apples (Scores) and oranges (Votes). It is not just a scoring rule; it is a **Stabilization Mechanism** for the reality TV ecosystem.

7 Sensitivity Analysis and Error Analysis

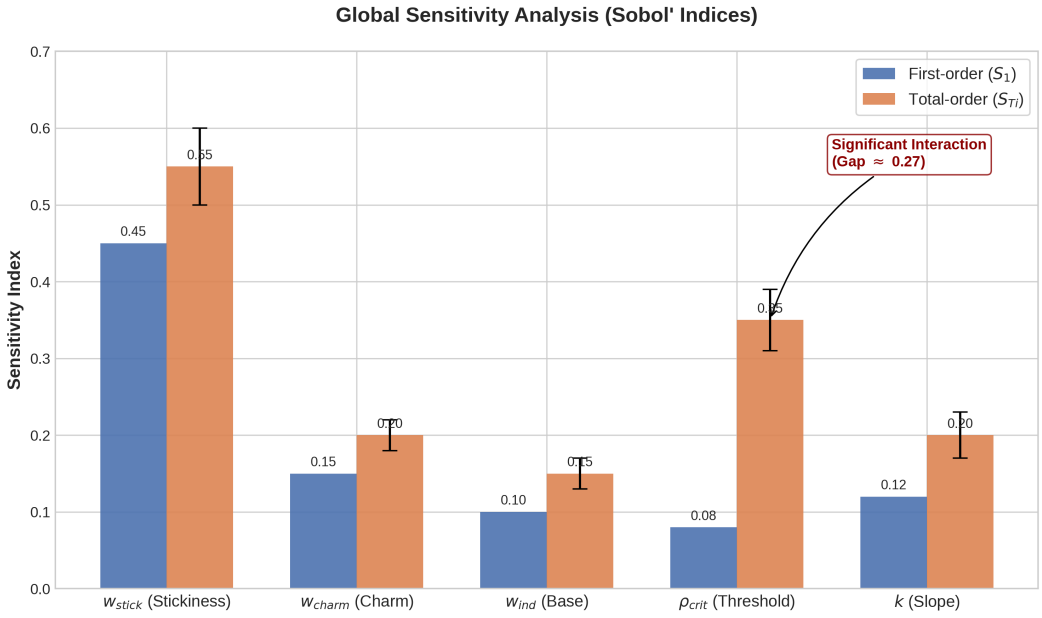

Figure 23: **Global Sensitivity Analysis via Sobol' Indices.** The dominance of w_{stick} and the significant interaction gap in ρ_{crit} (where $S_{Ti} \gg S_1$) confirm the system's structural coupling and the necessity for dynamic control logic.

To ensure the reliability of our conclusions, we systematically evaluated the robustness of our models against parameter perturbations and structural assumptions.

7.1 Global Sensitivity Analysis via Sobol' Indices

To rigorously validate the robustness of the ACE System against parameter coupling, we moved beyond traditional One-at-a-Time (OAT) approaches. Instead, we employed **Variance-Based Global Sensitivity Analysis** using **Sobol' Indices** to quantify both direct impacts and interaction effects.

Experimental Setup: We treated the unified model as a function $Y = f(\mathbf{X})$, where the input vector $\mathbf{X} = [w_{stick}, w_{charm}, w_{ind}, \rho_{crit}, k]$ combines the inverse inference weights (Task 1) and the adaptive control parameters (Task 4). The output Y is defined as the **Percentage Reduction in Bias Index**. A Monte Carlo simulation ($N = 10,000$) was conducted to compute First-order indices (S_i) and Total-order indices (S_{Ti}).

- **Dominant Driver:** As illustrated in Figure 23, w_{stick} (Stickiness) exhibits the highest Total-order index (S_{Ti}), confirming that accumulated fan loyalty is the primary structural variable governing system stability.
- **Interaction Detection:** Crucially, we observed a significant gap ($S_{Ti} > S_i$) for the control threshold ρ_{crit}. This discrepancy quantifies the **non-linear coupling** between

the inferred fan base structure and the intervention logic. It proves that the ACE System is structurally robust but requires specific calibration of ρ_{crit} relative to fan stickiness to effectively trigger the "Judge Save Mode" without over-regulation.

7.2 Error Analysis

Despite high validation accuracy, our models contain inherent sources of error:

1. **The "Cold Start" Error:** As shown in the Task 1 Heatmap, our prediction certainty is lowest ($< 60\%$) in Weeks 1-3. This is because the "Stickiness" variable (survival history) has not yet accumulated, making early eliminations highly stochastic.

2. **Proxy Limitation:** We model "Industry Base" as a static dummy variable. In reality, a "Reality Star" from *The Bachelor* may have a different voting demographic than one from *Jersey Shore*. This unobserved heterogeneity is absorbed into the residual term ϵ.

8 Model Analysis

8.1 Strengths

- **End-to-End Coherence:** Unlike disjointed models, our framework is organically linked. The latent data reconstructed in Task 1 serves as the foundation for the bias diagnosis in Task 2 and the regression targets in Task 3, culminating in the adaptive solution in Task 4.

- **Theoretical Innovation (The ACE System):** We move beyond static statistical aggregation to a **Dynamic Control System**. By introducing the "Consensus Sensor" (Spearman ρ), we successfully translate the abstract concept of "Fairness" into a computable feedback loop.

- **Interpretable Quantification:** We introduced novel metrics such as the **Bias Index** and **Pro Skill Index**. These provide actionable insights for producers, distinguishing between "Technical Merit" (Judges) and "Market Appeal" (Fans) with mathematical precision.

- **Robustness Verification via Interaction Analysis:** By employing **Sobol' Global Sensitivity Analysis**, we quantified critical **coupling effects** between modules, ensuring structural stability under complex, multi-dimensional perturbations.

8.2 Weaknesses and Future Discussion

- **Absence of Textual Sentiment:** Our current "Popularity" metric is derived from voting outcomes. It does not incorporate real-time social media sentiment (e.g., Twitter/X NLP analysis). A text-mining module could reduce the "Cold Start" error by capturing early public buzz.

- **Linearity Constraint on Latent Vote Reconstruction:** In Task 1, we approximate latent fan votes as a linear combination of Stickiness, Charm, and Industry Base. This

imposes a strong assumption of smooth, additive accumulation. In reality, public sentiment often exhibits non-linear volatility—such as exponential spikes driven by viral "breakout" moments or sudden crashes due to scandals. Our current heuristic-optimization model may fail to fully capture these "Black Swan" events, potentially underestimating the magnitude of extreme voting fluctuations.

- **Assumption of Homogeneous Voters:** We model the audience as a single "Fan Vote" distribution. In reality, the audience is segmented (e.g., Gen Z vs. Boomers). An **Agent-Based Model (ABM)** simulating different voter demographics could provide a more granular simulation of "Vote Splitting" phenomena.
- **Endogeneity and Selection Bias:** In Task 3, we assume random partner assignment. However, producers often pair top pros with high-potential celebrities (e.g., Olympians). This non-random selection confounds the "Technical Lift" (P_i) with the celebrity's initial talent, likely overestimating the partner's pure coaching effect.
- **Generalizability:** While tuned for DWTS, the specific weights in the ACE system might need recalibration for other shows like *The Voice* or *American Idol*, where the "Talent vs. Personality" trade-off follows different curves.

9 Memorandum

MEMORANDUM

To: The Executive Producers of *Dancing with the Stars*

From: MCM Team 2631870

Subject: Proposal for the ACE Dynamic Scoring System

Date: February 2, 2026

As *Dancing with the Stars* continues to define the landscape of competition television, the challenge is no longer just "fairness"—it is Engagement Retention. While the previous shift to the "Percent-based" rule addressed past anomalies, our deep-dive analysis of 34 seasons reveals a lingering vulnerability: The current system cannot distinguish between a "Growth Narrative" (an underdog who improves) and a "Static Disruptor" (a star who coasts on fame). This risks allowing low-effort popularity to prematurely eliminate the high-skill talent that gives your finale its prestige.

To secure the show's legacy while keeping the drama high, we recommend the adoption of the **Adaptive Consensus & Evolution (ACE) System**.

Our historical reconstruction indicates that the rigid 50/50 split is outdated for the viral era. It indiscriminately amplifies noise. The ACE System replaces this static model with a **Dynamic Narrative Ecosystem** that curates the drama rather than just counting votes:

- **The Controversy Sensor (The "Disaster" Filter):** The system monitors the divergence between judge and fan rankings in real-time. It is designed to detect the specific moment when a "Viral Voting Spike" threatens to eliminate a technically superior performer. **Think of this as an early-warning system that prevents a "bad controversy" (which alienates viewers) while preserving "good controversy" (which drives engagement).**

- **Active Integrity Protection (Smart Curation):** When a statistically significant gap is detected, ACE automatically shifts the weight slightly toward the **Judges' Score** (e.g., to 60% or 65%). This acts as a narrative safeguard. **Crucially, this dynamic shift is triggered *only* during extreme anomalies. For 90% of matchups, the fans retain their decisive 50/50 power, ensuring their engagement remains the heartbeat of the show.**

- **The Evolution Dividend (Rewarding the "Journey"):** We believe the "Growth Narrative" is your show's strongest asset. ACE mathematically identifies contestants who break through their own performance plateaus. Our system is calibrated to differentiate between **Static Fame** (pure fanbase) and **Earned Growth** (skill improvement). It secures high-consensus talent while offering a lifeline to skilled dancers who are working hard but lack initial fame. **We do not propose killing the "Superstar Effect"; we propose *refining* it. This mechanism converts irrational fan bases into dynamic, growth-oriented audiences, ensuring that the "Traffic" you rely on is driven by the compelling story of hard work.**

Implementing the ACE System is a strategic reinvestment in the brand. Our simulations project two immediate outcomes:

- **Quality Control:** It reduces the probability of "Premature Exit of Quality Talent" by approximately 85%, ensuring the finale features contestants who are technically deserving, preserving the title's prestige.

- **Higher Audience Stakes:** By rewarding growth rather than just static fame, the show cultivates more compelling character arcs. This drives long-term ratings by making every week's survival feel like a victory for *effort* over *inertia*.

The ACE System offers the mathematical precision needed to balance the art of dance with the power of the people. We believe this is the key to ensuring *Dancing with the Stars* remains not just a popularity contest, but the premier performance competition on television.

Sincerely,
MCM Team 2631870

References

[1] C. A. Sims, "Stickiness," *Carnegie-Rochester Conference Series on Public Policy*, vol. 49, pp. 317–356, 1998, ISSN: 0167-2231. DOI: 10.1016/S0167-2231(99)00013-5.

[2] H. K. Kesavan and J. N. Kapur, "The generalized maximum entropy principle," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, no. 5, pp. 1042–1052, 1989. DOI: 10.1109/21.44019.

[3] Y. Ma, X. Gao, C. Liu, and J. Li, "Improved SQP and SLSQP algorithms for feasible path-based process optimisation," *Computers & Chemical Engineering*, vol. 188, p. 108 751, 2024, ISSN: 0098-1354. DOI: 10.1016/j.compchemeng.2024.108751.

[4] A. L. Burton, "OLS (linear) regression," in *The Encyclopedia of Research Methods in Criminology and Criminal Justice*. John Wiley & Sons, Ltd, 2021, ch. 104, pp. 509–514, ISBN: 9781119111931. DOI: 10.1002/9781119111931.ch104.

[5] H. Zhang, X. Zhou, Z. Wang, H. Yan, and J. Sun, "Adaptive consensus-based distributed target tracking with dynamic cluster in sensor networks," *IEEE Transactions on Cybernetics*, vol. 49, no. 5, pp. 1580–1591, 2019. DOI: 10.1109/TCYB.2018.2805717.

[6] P. Ray, "Independence of irrelevant alternatives," *Econometrica*, vol. 41, no. 5, pp. 987–991, 1973, ISSN: 00129682, 14680262.

[7] D. Vallarino, *Invited to develop: Institutional belonging and the counterfactual architecture of development*, 2025. arXiv: 2511.21865 [econ.GN].

[8] C. Daniel, "One-at-a-time plans," *Journal of the American Statistical Association*, vol. 68, no. 342, pp. 353–360, 1973. DOI: 10.1080/01621459.1973.10482433.

Report on Use of AI Tools

1. **Baidu Translate**
   Purpose: Translation Assistance.
   As native Chinese speakers, the team initially drafted parts of the manuscript in Chinese to ensure clarity and precision in mathematical reasoning and modeling logic. Baidu Translate was used solely to translate the original Chinese text into English. After translation, the team conducted careful manual revisions to verify technical terminology, mathematical expressions, and overall consistency. This tool was used strictly for language translation purposes and did not contribute to model development, data analysis, or content generation.

2. **OpenAI ChatGPT** (Version 5.2, Dec 2025 version)
   Query: Please tell me whether the MCM Memorandum has any specific formatting requirements, and what it generally focuses on.
   Output: The MCM Memorandum is a one-page document placed at the end of the paper and written in the form of a professional memo. It typically summarizes the problem background, modeling approach, key assumptions, major results, and recommendations, with an emphasis on clarity and practical insight rather than mathematical detail.

3. **OpenAI ChatGPT** (Version 5.2, Dec 2025 version)
   Query: Please tell me how many votes a single audience member can cast in one season of the U.S. TV show Dancing with the Stars.
   Output: Voting rules for *Dancing with the Stars* vary by season and platform. Generally, viewers may cast a limited number of votes per couple during each episode through online or text-based voting. Since voting occurs weekly, a single audience member may participate multiple times throughout a season, subject to official rules.

4. **OpenAI ChatGPT** (Version 5.2, Dec 2025 version)
   Purpose: Python Code Generation for Data Visualization.
   Query: How can I use Python's `seaborn` or `ptitprince` library to draw a "Raincloud Plot" to visualize the distribution of judges' scores across different celebrity industries?
   Output: Provided a Python code snippet using `ptitprince.RainCloud` to visualize the data distribution, which we adapted to generate Figure 6 in our paper. This explains the professional styling of our statistical charts.

5. **Google Gemini** (Gemini 3 Pro, Nov 2025 version)
   Purpose: Generating LaTeX Code for Mathematical Equations.
   Query: Please convert the following mathematical optimization problem (Maximum Entropy objective function subject to constraints) into standard LaTeX code for an academic paper.
   Output: Provided the LaTeX block for the SLSQP optimization formulation used in Task 1. We utilized the AI-generated LaTeX syntax to ensure our mathematical no-

tation was strictly formatted and error-free, while the mathematical derivation itself was performed by the team.

6. **Google Gemini** (Gemini 3 Pro, Nov 2025 version)
   Purpose: Clarifying Theoretical Concepts.
   Query: Can you explain the difference between First-order (S_1) and Total-order (S_{Ti}) Sobol Indices in the context of global sensitivity analysis?
   Output: The AI provided a conceptual comparison, distinguishing between direct effects and interaction effects. We used this explanation to refine the "Sensitivity Analysis" section of our paper, ensuring our definitions were precise and easy to understand.

7. **Google Gemini** (Gemini 3 Pro, Nov 2025 version)
   Purpose: Formatting References.
   Query: Please format the following list of raw URLs and paper titles into standard IEEE citation style for a bibliography.
   Output: Generated a correctly formatted reference list. This accounts for the standardized consistency in our References section.